

XML Metadata for Accessing Heterogeneous Legal Databases

Virpi **Lyytikäinen** <lyviau@cc.jyu.fi>

Pasi T. **Tiitinen** <pti@cc.jyu.fi>

Airi **Salminen** <asalminen@db.uwaterloo.ca>

Abstract

Legal information in Europe is scattered in numerous heterogeneous databases. The data in the databases is structured, organized and classified in various ways, the contents are written in different languages, and the retrieval techniques vary. Providing integrated access to the databases would serve both legal experts and laymen. Issues related to the Web access of European legal databases were studied in the EULEGIS project. Requirements for the integrated service were investigated and a prototype system was implemented. The implementation was based on the idea of rich metadata. An XML-based model for the metadata was developed and implemented. The model included data about legal processes, organizational actors in the processes, types of documents created in the processes, and databases providing access to the documents of the types. An important subset of the metadata was visualized in the user interface graphically. The paper describes the metadata model and how the metadata is used in the user interface.

1. Introduction

Legal documents in Europe are created in different legal systems and stored in numerous heterogeneous databases. Documents in the databases are structured, organized and classified in various ways, their contents are written in different languages, and their retrieval techniques vary. One database often includes documents from one legal system. Economic integration in Europe and the globalization more generally has caused a situation where decisions at organizations and also private decision-making requires legal information originating from different

legal systems. In spite that a great deal of the documents are accessible over the Internet, the heterogeneity of the databases and their retrieval techniques causes difficulties in information access.

Issues related to the Web access of European legal databases were studied in the European User Views to Legislative Information in Structured Form (EULEGIS) project funded by the European Union Telematics Application Programme. The two-year project ended in May 2000. Requirements for an integrated legal information service were investigated and a prototype system was implemented. The implementation was based on the idea of rich metadata. The core of the EULEGIS system is a relational metadata database. The structure of the metadata is described by Extensible Markup Language (XML) Document Type Definition (DTD) [BPS 2000], and XML is used as the format in the data exchange between EULEGIS modules. During the prototype implementation and data collection, XML was also used as the exchange format between people collecting the data.

Association of metadata with resources is well-known, traditional practice in libraries. Lately special metadata models and schemes have been developed for the management of Internet resources [GIL 1998]. Resource Description Framework (RDF) is a general metadata model recommended by World Wide Web Consortium (W3C) [LAS 1999]. RDF descriptions are statements about resources identified by Uniform Resource Identifier (URI)s, and the statements can be written using XML syntax. In Dublin Core Initiative [BMB 2000] a vocabulary for associating a set of metadata elements with electronic resources is introduced. The Dublin Core metadata can be described by the RDF syntax. An example of a metadata scheme on a specific domain is the IMS metadata model for learning resources [IMS 2000]. In the EULEGIS project a special metadata model for legal information was developed. While in the above mentioned metadata models and schemes metadata is typically data about documents, the EULEGIS metadata is data about the legal systems where documents are created together with data about databases where the documents are available.

2. Modularization of the Metadata

As discussed above, the [EULEGIS](#) metadata database does not contain data about legal documents directly. Instead, it contains data about legal systems and data about legal databases. Data about legal systems is used to show to the users the context in which the documents in databases were created. The interviews carried in the [EULEGIS](#) project showed that one cause of problems in legal information access derives from the differences of European legal systems. For being able to access data originating from different legal systems the users need to have at least some kind of understanding of the systems.

A *legal system* can be defined as a set of legal rules governing, for example, a state, a group of states, an international organization, a region, or a city. In Europe, examples of legal systems include

- the European Union,
- a single member state (e.g. Finland, Sweden, United Kingdom or France),
- a single region (e.g. the German state of Bavaria), or
- a single municipality.

Although all legal systems have some resemblance to each other, they are all different in detail (see e.g. [[GLM 1995](#)]). A major difference in the legal systems of different countries is in the role of regions. In some countries (e.g., Germany, Austria and Belgium), the legislative power has been decentralized to several power centers, which have their own legal systems. On the other hand, there are countries only with minor legislative power at regional level. Certain amount of regulative power has often also been left, e.g., to provinces or municipalities.

In spite that legal systems vary, they all have the following characteristics:

- They have *actors* (such as the Parliament or Ministry of Agriculture).
- Their rules, the creation of the rules, and the application of the rules is documented in *legal documents* of different document types (e.g. Statutes or Government Bills).

- Each implements its own *process*(creation of legal sources).

The **EULEGIS** metadata model is based on these common characteristics. **Figure 1** depicts the modularization of the **EULEGIS** metadata. The arrows show how the modules are linked to each other. There are four kinds of modules:

1. **DATABASE**. Describes a legal database.
2. **PROCESS**. Describes the legal process of a legal system.
3. **ACTORS**. Describes the organizations and persons involved in the legal process of a legal system.
4. **DOCUMENTS**. Describes the document types created in the legal process of a legal system.

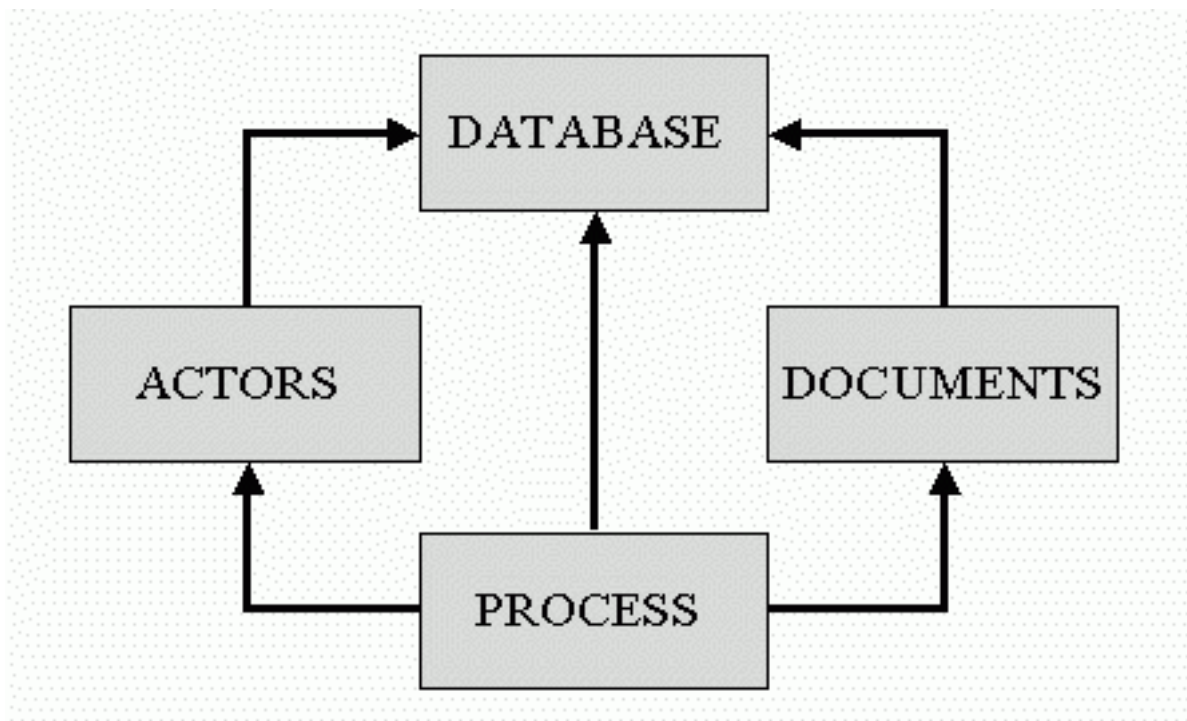


Figure 1. Metadata model

The metadata database contains one **DATABASE** description for each legal database, and one **PROCESS**, **ACTORS**, and **DOCUMENTS** description for each

legal system. In the following section the [XML DTDs](#) of the four modules are described. The graphical forms shown in figures have been produced using the Near & Far Designer 3.0. The data created based on these [DTDs](#) is interconnected according to the arrows shown in [Figure 1](#).

3. The Four Metadata Modules

3.1. Data about Legal Databases

The metadata concerning databases serve two main functions in the [EULEGIS](#) system. Firstly, they enable a unified interface for querying all different databases to be formed. Secondly, they aid in unifying the result of a query so that it looks approximately similar to the user regardless of the original database.

The participating legal databases share the common feature that their *query interfaces* are based on the use of Web forms. There can be one or more different forms for one database. Each query interface in turn contains one or several *query fields*, some of which the user is supposed to fill, when presenting the query. A subset of the query fields can be hidden from the user in the query interface, but still they contain valuable information from the database management's point of view. The metadata description contains information about each query interface of the database and about each of the query fields either visible or hidden to the user. Also information about allowed operators, for example, Boolean operators or wild characters, and about their names are included in the metadata.

To make the [EULEGIS](#) system multilingual all informative content in the metadata database was described in several languages. [Figure 2](#) shows the outermost elements of a [DTD](#) fragment that was used for describing a query interface. The element '*interfacename*' contains the name of the original query interface (such as a page for querying decisions of the Supreme Court in a Finnish legal database, Finlex), the element '*databasename*' the name of the database (such as Jurifrance or CELEX), the element '*interfaceurl*' provides the Uniform Resource Locator ([URL](#)) of the original query interface, and finally the element '*databaseurl*' contains the [URL](#) of the home page of the database. Attributes of the element '*interface*' contain information e.g. about the languages of the documents in the database.

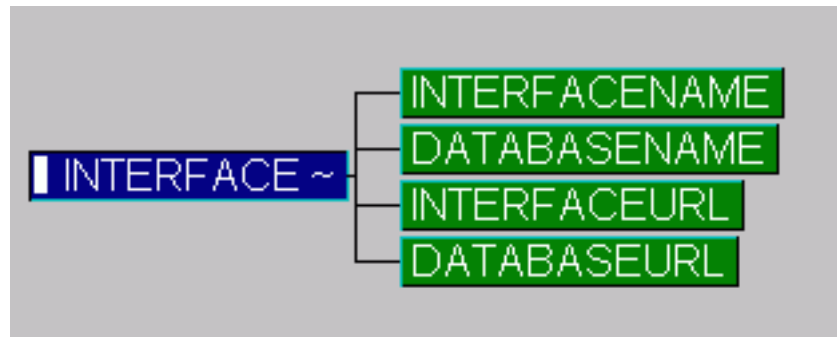


Figure 2. DTD Fragment for describing a query interface

Information about the search fields of a certain original query interface is described by the [DTD](#) fragment presented in [Figure 3](#). The description of search fields can be hierarchical thus allowing the grouping of fields for further handling as a whole. This is useful, for example, in the case of dates, which are often given in original interfaces by using separate input fields for day, month and year. Information about a specific field in a query interface is described by the 'fielddescr' element. The 'name' element contains the name of the HyperText Markup Language ([HTML](#)) form field as it appears in the query interface. The field labels, which are shown to the users, are provided in a 'label' element. The 'description' element may be used to provide more information about the purpose of the field.

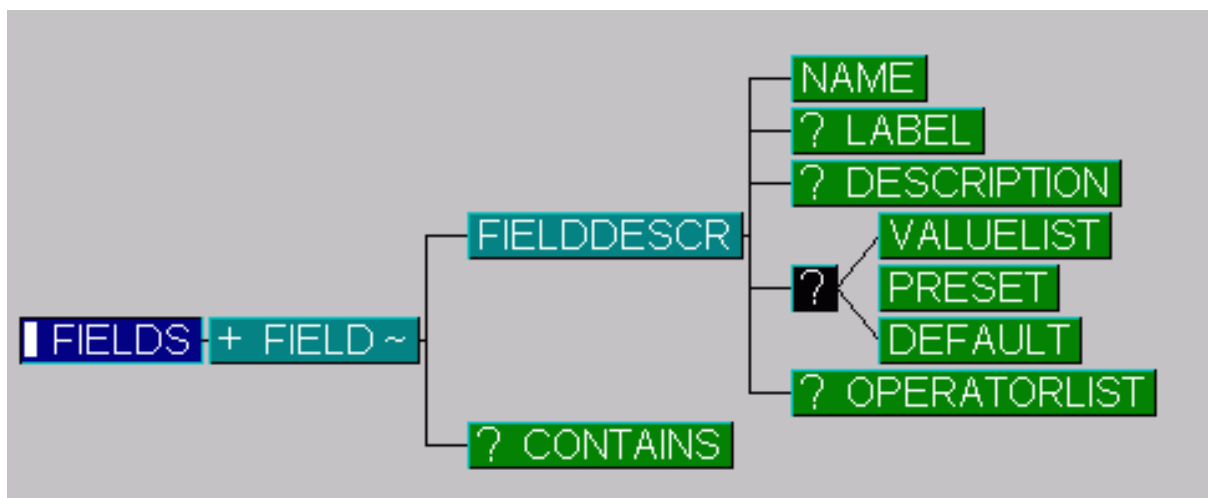


Figure 3. DTD Fragment for describing query fields

In some cases, there can be a list of options or allowed values related to a field, which may be defined in the optional '*valuelist*' element. It is also possible to define a preset or default value. In many cases logical operators, such as AND, OR, NOT, and wildcard character operators are allowed in the original interface. These are described by the '*operatorlist*' element. Hierarchical representation of fields has been accomplished by using the '*contains*' element, which contains references to other fields. An attribute of the '*field*' element contains also information about the type of the field. Possible types include text, number, Boolean, date and date span.

A major task in the creation of metadata about the databases was the collection of the data from the database providers, for example, concerning how the search fields of different query interfaces relate to each other. In the original query interfaces there are fields that can be used, for example, for free text search, for limiting a query to the titles of the documents, or for searching for documents by their publication date. The field labels, which are presented to the users, however, may vary or may even be in different languages in different query interfaces. In order to provide a unified query interface all different query fields having the same semantic purpose (for example "title search") should be provided with a common name. In the [EULEGIS](#) system these unified query fields are then called [EULEGIS](#) search fields. The [EULEGIS](#) search fields enable searching of multiple databases by a single query. The mappings between the [EULEGIS](#) search fields and the fields of the original query interfaces are described in the metadata using the [XML Linking Language](#) [[DMO 2000](#)].

3.2. Data about Legal Actors

The term *legal actor* refers to the organizations or persons involved in the process of producing legal documents in a legal system. The actors can form actor groups as, for example, in the Parliament there can be special committees and other groups participating the legislative process.

The [DTD](#) for describing legal actors is depicted in [Figure 4](#). The root element of the [DTD](#) is called '*actors*'. The '*actors*' element consists of actor groups called '*actorgrp*'s in the [DTD](#). An actor group again consists of actor groups or actors. Each actor group and actor must have a '*name*' and a unique identifier defined as an '*id*' attribute. The name can be given in multiple languages. By the element '*role*' tasks of

an actor group or actor can be described shortly. Like the names, also the roles can be expressed in multiple languages. For expressing additional information concerning actor groups, actors and their roles in the legal system, the DTD includes an optional element 'description'. The content of the 'description' element can either be a mixture of paragraphs and titles or a link to an external file. The 'targetdb' element defines a link to a database containing documents created by the actor group in question. All links in the EULEGIS DTDs are defined using the XML Linking Language (XLink) language [DMO 2000].

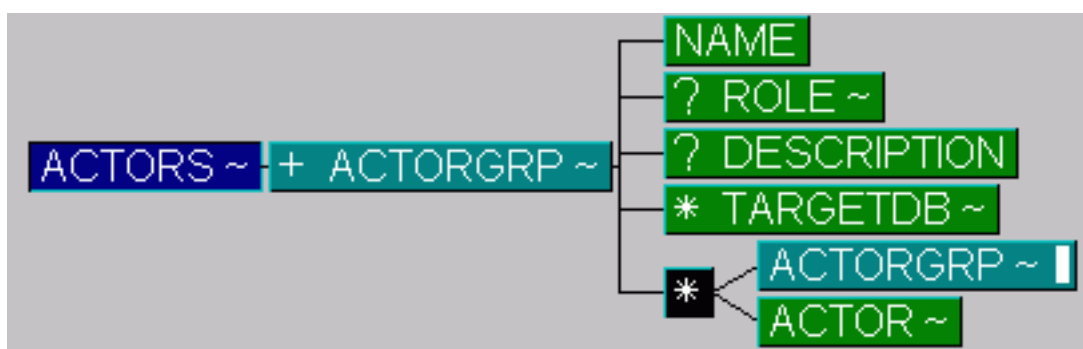


Figure 4. DTD for describing legal actors

3.3. Data about Document Types

For each legal system there is a description of the legal document types of the system. Also, the term *legal information sources* is used to refer to the types. The grouping of document types expresses their hierarchic containment relationships. Some document types can also have precedence over other in the legal system; this is also indicated in the description.

The DTD for describing the document types of a legal system is depicted in Figure 5. The root element of the DTD is called 'docs'. The 'docs' element consists of title called 'name' (e.g. 'Legal information sources related to Finnish legal system') and of one or more 'docgroup' elements representing the document groups in the legal system. A document group again contains other 'docgroup' elements or a reference to document groups in other legal systems ('doclink' element). Each document group must have a 'name' and a unique identifier defined by an 'id' attribute. The name can be given in multiple languages. Hierarchic relationships between document groups (e.g. Constitution having precedence over Act) are defined by an attribute. For

expressing additional information concerning document groups, the document type definition includes an optional element '*description*'. The content of the '*description*' element can either be a mixture of paragraphs and titles or a link to an external file. The '*targetdb*' element is a link to a database containing the documents belonging to the document group in question.

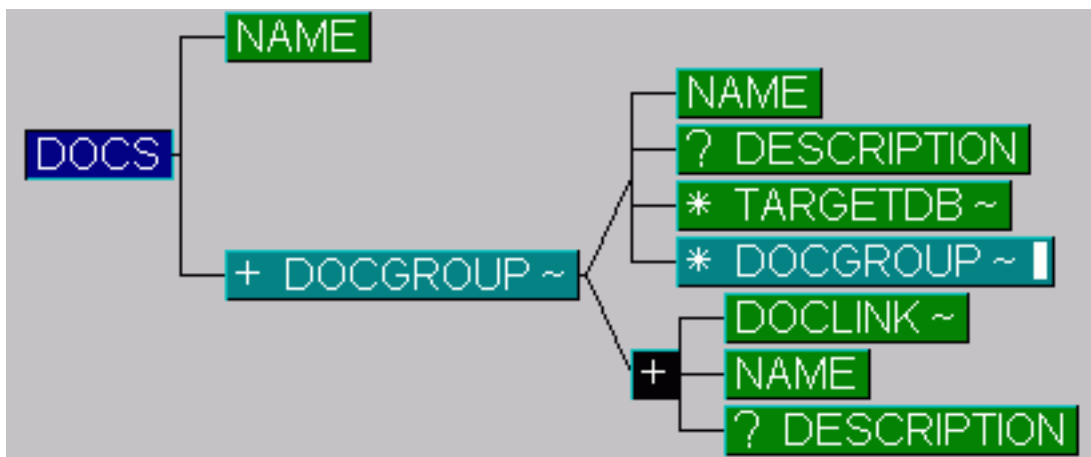


Figure 5. DTD for describing legal information sources

3.4. Data about Legal Processes

A *legal process* consists of sequential, parallel or alternative activities. Each activity is performed by some legal actor, and as a product of the activity there can be documents of specific types.

The DTD for the data about a legal process is shown in Figure 6. The root element of the DTD is called '*pros*'. The '*pros*' element consists of the name (*name*) of the legal process in question (e.g. 'Creation of legal sources in Finland'), activities (*activity*), and connectors (*connect*).

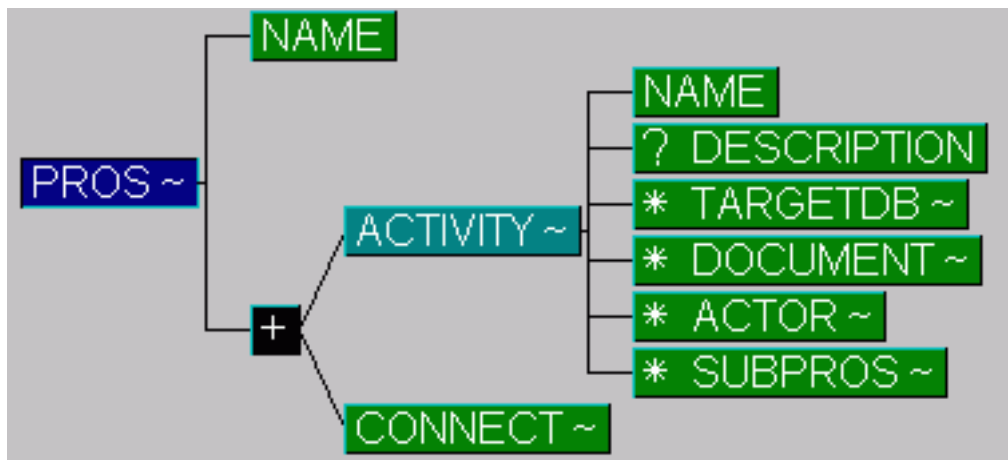


Figure 6. DTD for describing legal processes

Each 'activity' element has a 'name', an optional element 'description' for additional information concerning the activity, links to documents produced during the activity, and links to actors performing the activity. The activities following the current activity in the process are identified by an attribute. The content of the 'description' element can either be a mixture of paragraphs and titles or a link to an external file. The 'targetdb' element defines a link to a database containing documents created in the activity in question. To establish a link towards a document produced during the activity, an identification of that document type is needed. The reference to the document is placed to element named 'document'. Similarly, to establish a link towards an actor performing the activity, an identification of that actor is needed. With 'subpros' element links to possible subprocesses are made. Connectors are either AND connectors or OR connectors. They are needed to define parallel or alternative activities in a process. The type of the connector is indicated by an attribute.

4. Visualization of the Metadata

The data about the legal actors, document types (or legal information sources), and processes of a legal system is displayed to the users in a graphical user interface. Since each of the three metadata descriptions in a way represents a special kind of view of the legal system, they are called *actor view*, *legal information source view*, and *process view*, respectively, in the user interface. The graphical representations of the metadata are actually graphical data models originally developed for document

analysis purposes [[SAL 2000](#)].

4.1. Actor View

The actor view shows to the user the most significant actors that are involved in the legal process of a legal system. The actors and actor groups are depicted by rectangles and the hierarchic relationships of them are indicated by nesting. The actors belonging to an actor group may be part of a larger organization (e.g. Parliament), or they may otherwise have similar roles in the process (e.g. different courts). In the bottom part of the window the tasks of the actors in the process are shortly described. A sample actor view is shown in [Figure 7](#).

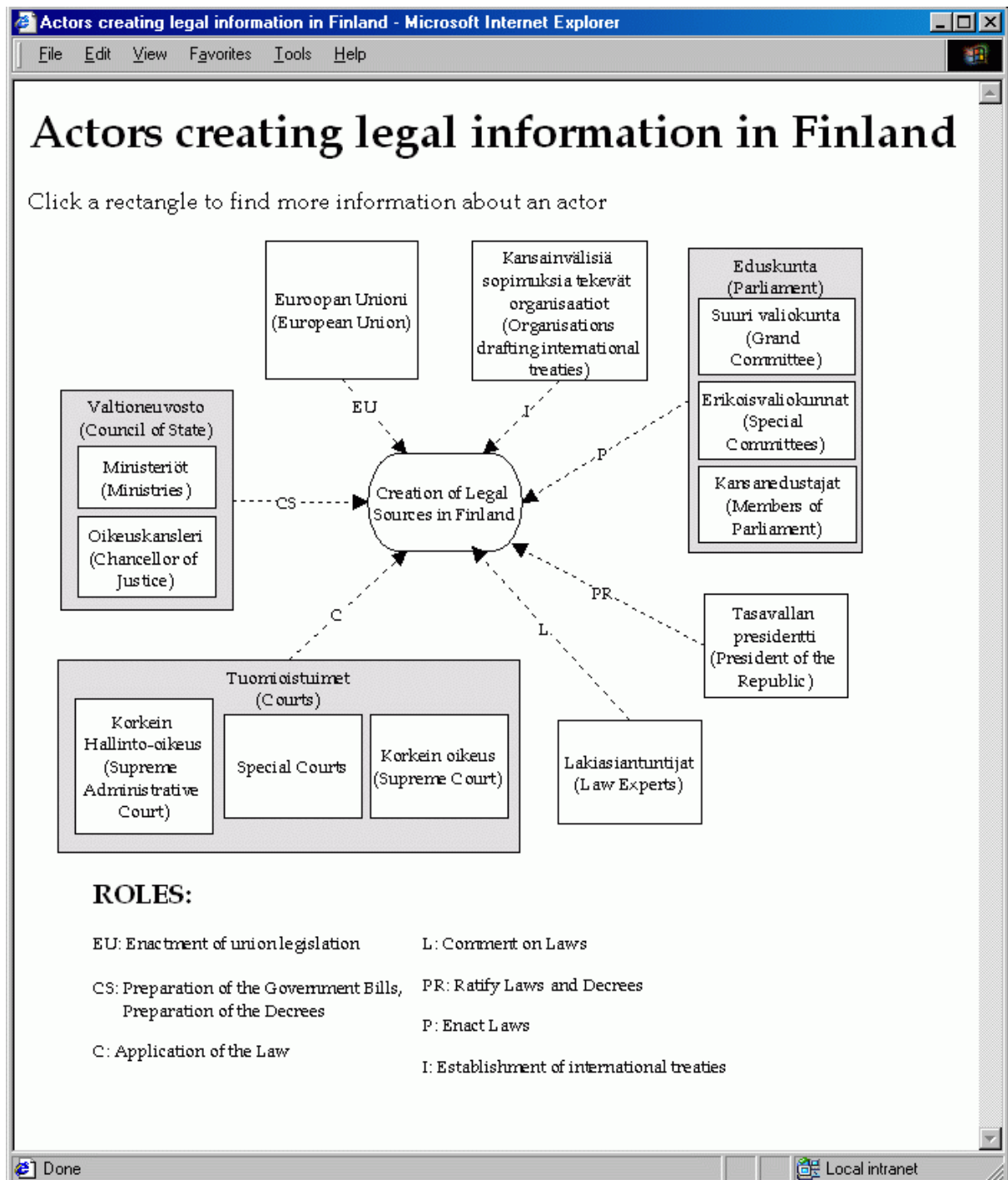


Figure 7. The actor view to the Finnish legal system

4.2. Legal Information Source View

The legal information source view shows to the user the most significant legal document types in a legal system and their grouping. [Figure 8](#) describes the legal information sources of the Finnish legal system as an example. The document types and document groups are indicated by rectangles. The Finnish legal documents are categorized into five groups that exist in almost every country: preparatory works, normative documents, court judgements, legal literature, and miscellaneous parliamentary documents (including, for example parliamentary questions and minutes of the plenary sessions). The document types connected by vertical lines represent a set of documents in a hierarchic relationship in the legal system. For example, in the figure such types are constitution, act, decree, and administrative act.

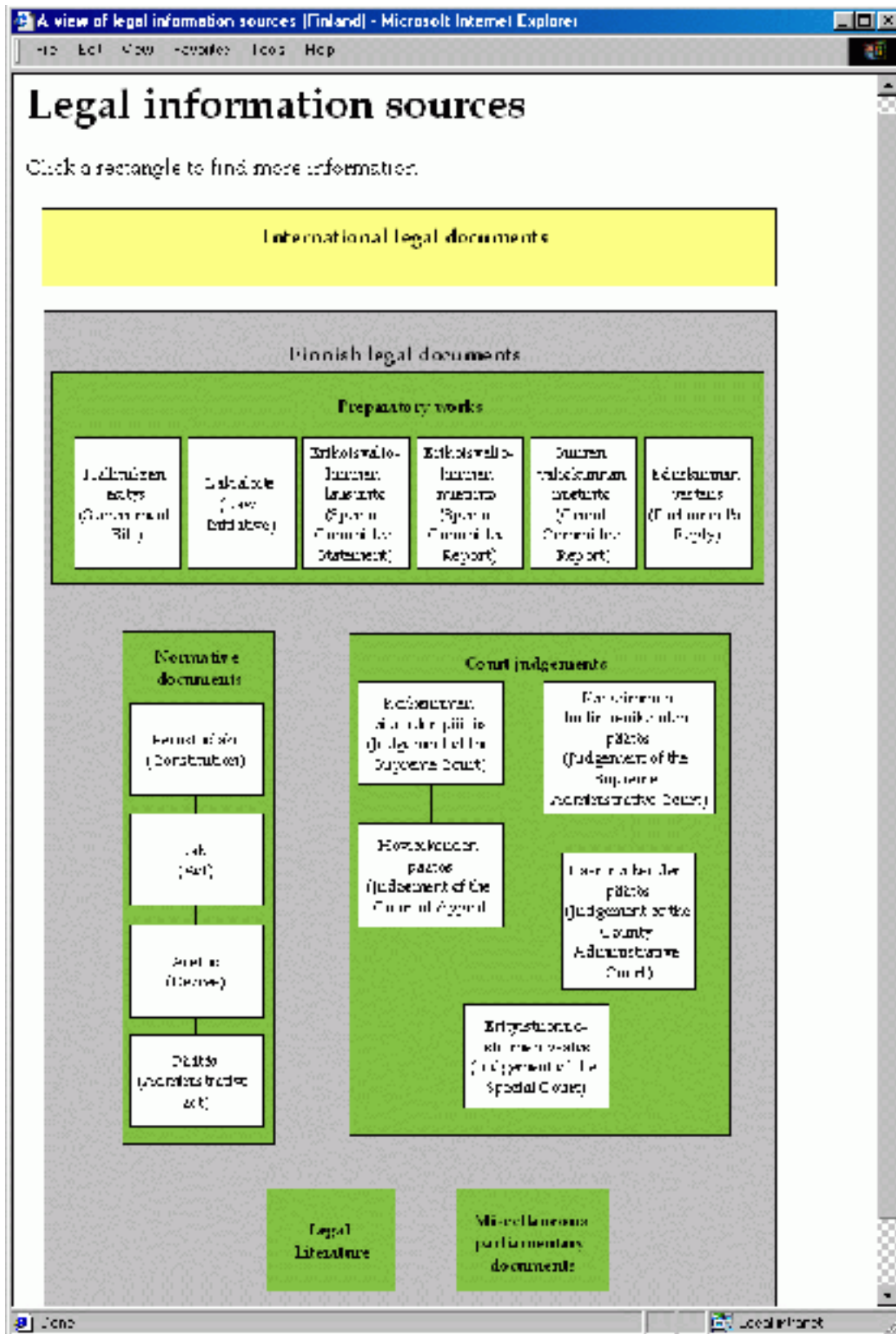


Figure 8. Legal information source view of Finland

4.3. Process View

The process view gives to the user a graphical overview of the most significant activities in a legal system. It contains information about the order of activities, about the actors responsible for performing the activities, and about the documents created in those activities. As an example of the process view the Finnish national legal system is presented in [Figure 9](#). The activities are depicted by circles. Each circle shows both the name of the activity and the actor(s) performing the activity. The order of the activities is expressed by solid arrows. Alternative activities are indicated by a hollow dot and parallel activities by a black dot. The documents created in the activity are shown by a dashed arrow labeled with the type of the document.

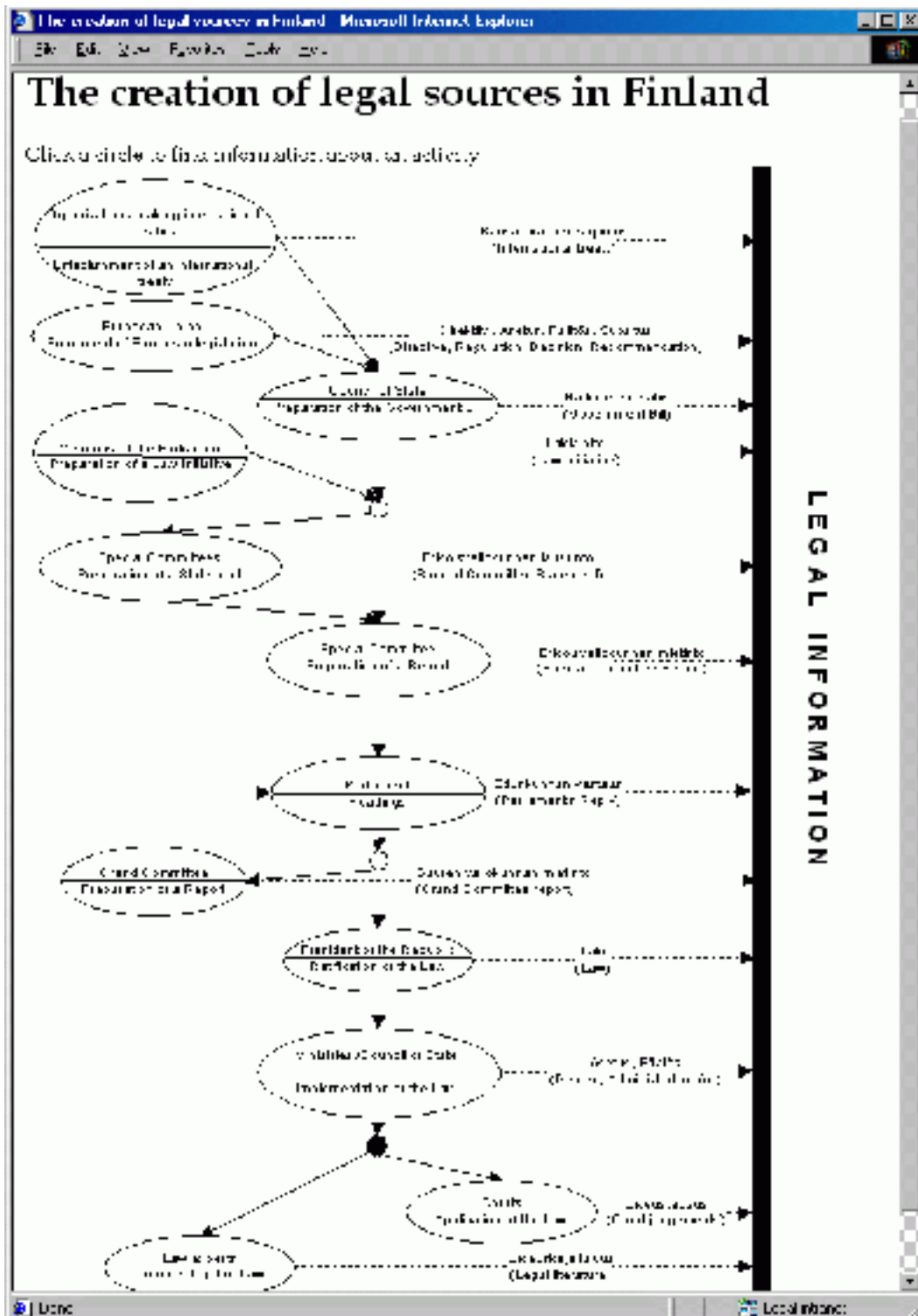


Figure 9. Process view to the Finnish national legal system

5. Accessing Legal Data

There are two basic approaches for accessing legal information in the [EULEGIS](#) system: by choosing one or more databases directly or by using the graphical views. In both approaches the user chooses the legal system(s) whose legislation she or he is interested in. The [EULEGIS](#) prototype was connected to legal databases in Finland, France, Portugal, Sweden, and the United Kingdom. In addition to the national databases, the user had access to the CELEX database containing legal documents created in the different organizations of the European Union. The legal systems described in the prototype and supported by the graphical views consisted of the corresponding national legal systems and the European Union legal system.

The graphical views provide two kinds of functionalities. On the one hand they provide general information about the legislative processes, actors and document types in a certain legal system. By clicking a symbol representing a specific activity, actor, or document type, the user is able to see a textual description of the object in interest on a separate window. The second way of using a graphical view is as an aid in choosing appropriate [EULEGIS](#) search forms for retrieving documents. By clicking any of the objects visible on the graphical representation the user can get a search form, which enables him or her to query the documents created by the selected actor, created in the selected phase of the legislative process, or belonging to the selected group or type of documents. The user needs not to know in formulating the query in which database(s) the documents of interest are available.

After the user has submitted a query, the database provides a list of matching documents. Since each of the original databases has its own way to display the results of the query, the [EULEGIS](#) metadata is used to allow the representation of the resulted hitlists in a common format. The original hitlists were analyzed in the project and based on that analysis a simple [XML DTD](#) was designed for a unified hitlist format. The [HTML](#) formatted hitlists resulting from a query are converted to conform to this [DTD](#), and back to [HTML](#) to be shown in relatively consistent [EULEGIS](#) layout to the user. The conversions are implemented using LotusXSL, which is an Extensible Stylesheet Language Transformations ([XSLT](#)) processor for transforming [XML](#) documents into [HTML](#), text, or other [XML](#) document types. The user can select interesting documents from the hitlist, which are then accessed from the databases.

These documents are also converted via [XML](#)middle format to [HTML](#) . The user can see the documents in a layout made as uniform as possible for different document types and for documents accessed from different databases.

6. Conclusion

One of the motivations in initiating the [EULEGIS](#) project was to explore the capabilities of the Standard Generalized Markup Language ([SGML](#))/[XML](#) technology to improve the accessibility of European legal information. In the begin of the project the development of a common [DTD](#) for European legal documents was discussed. In spite that some work was already done on this area earlier [[MAS 1998](#)], the project participants early agreed that the development of the [DTD](#) is too tedious task in a two-year project. Instead other ways to utilize the [SGML/XML](#) technology were tested. In the paper we described how [XML](#) definition capability was used to describe the metadata model for the [EULEGIS](#) retrieval tool. The use of [XML](#) as a definition technique directly facilitated the use of the [XML](#) format for exchanging the metadata between different software modules. Furthermore, the [XML](#) format was also found useful in exchanging data between people at the time of the metadata collection for the prototype system.

Building a uniform way for information rendering is one of the difficult challenges in building an integrated access to a set of legal databases. Although the hitlists and retrieved documents in the legal databases connected to the [EULEGIS](#) prototype were originally in [HTML](#) format, the way they were marked up varied greatly. Some database providers rendered their hitlists and documents only with a minimum amount of [HTML](#) tags, while some others used markup with layout oriented tagging, for example, by using tables for defining the layout of the Web page. The use of Extensible HyperText Markup Language ([XHTML](#)) would have simplified the task of unifying the query results from different databases, because [XHTML](#) requires complete tagging and recommends the use of CSS as a way to define the layout of Web pages.

Building an integrated access to heterogeneous European legal databases is a complex task where the technological problems are closely related to the complexity of the legal domain and to the fact that there are tens of languages in use in Europe. The feedback of the end users of the [EULEGIS](#) prototype however encourages

continuing the work on the area. The capabilities of the prototype were widely demonstrated during the project both to legal expert and laymen. The visualization of the legal systems was considered as an important means for making the legal processes more understandable. XML in the future implementation will be useful technology for the data exchange between different components of the distributed system.

Acknowledgements

Nine full partners and five associate partners from nine European countries worked together as a consortium in the EULEGIS project. People from all partner organizations have contributed to the development of the solutions described in the paper. The EULEGIS system was implemented by Atos, Indra, Katholieke Universiteit Leuven, and TietoEnator. We gratefully acknowledge the financial support provided by the Telematics Application Programme of the European Commission, the Academy of Finland (under Project 48989), and the National Technology Agency of Finland as well as the other funding partners in the inSGML project.

Bibliography

[BMB 2000] Beckett, Dave, Miller, Eric, and Dan Brickley, 2000. *An XML Encoding of Simple Dublin Core Metadata*. Dublin Core Metadata Initiative Proposed Recommendation. URL: <http://dublincore.org/documents/2000/11/dcmes-xml/> (Retrieved March 6, 2001).

[BPS 2000] Bray, Tim, Paoli, Jean., Sperberg-McQueen, C. M., and Eve Maler, (Eds.), 2000. *Extensible Markup Language (XML) 1.0 (Second Edition)*. W3C Recommendation 6 October 2000. URL: <http://www.w3.org/TR/2000/REC-xml-20001006/> (Retrieved March 6, 2001).

[DMO 2000] DeRose, Steve, Maler, Eve, and David Orchard, (Eds.), 2000. *XML Linking Language (XLink) Version 1.0*. W3C Proposed Recommendation, 20 December 2000. URL: <http://www.w3.org/TR/2000/PR-xlink-20001220/> (Retrieved March 6, 2001).

[GLM 1995] Gallagher, Michael, Laver, Michael, and Peter Mair, 1995. *Representative government in modern Europe*. McGraw-Hill. New York.

[GIL 1998] Gilliland-Swetland, Anne. J., 1998. Defining metadata. In M. Baca (Ed.) *Introduction to Metadata*, (pp. 1-8). Los Angeles: Getty Information Institute. Also available on WWW: URL: http://www.getty.edu/research/institute/standards/intrometadata/2_articles/index.html (Retrieved March 6, 2001).

[IMS 2000] IMS, 2000. *IMS Learning Resource Meta-data Information Model*. Version 1.1 - Final Specification. URL: <http://www.imsproject.org/metadata/mdinfov1p1.html> (Retrieved March 7, 2001).

[LAS 1999] Lassila, Ora and Ralph R. Swick, 1999. *Resource Description Framework (RDF) Model and Syntax Specification*. W3C Recommendation 22 February 1999. URL: <http://www.w3.org/TR/REC-rdf-syntax/> (Retrieved March, 8, 2001).

[MAS 1998] Magnusson Sjöberg, Cecilia. (1998). *Critical Factors in Legal Document Management*. Stockholm: Jure.

[SAL 2000] Salminen, Airi, 2000. Methodology for document analysis, in A. Kent and C. Hall, (Eds.), *Encyclopedia of Library and Information Science*, 67, Marcel Dekker, Inc, New York, 299-320.

Glossary

DTD	Document Type Definition
EULEGIS	European User Views to Legislative Information in Structured Form
HTML	HyperText Markup Language
RDF	Resource Description Framework
SGML	Standard Generalized Markup Language

URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XHTML	Extensible HyperText Markup Language
XLink	XML Linking Language
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

Biography

Virpi **Lyytikäinen**

Researcher

University of Jyväskylä

Department of Computer Science and Information Systems

Finland

Email: lyviau@cc.jyu.fi

Virpi Lyytikäinen — Virpi Lyytikäinen has been doing research related to structured SGML/XML documents at the Department of Computer Science and Information Systems in the University of Jyväskylä since 1996. At the moment she works in a project called inSGML, which is developing, testing and customizing methods for the SGML/XML standardization process especially for industrial purposes. The work is related to her Doctoral Thesis, whose subject is methods for SGML standardization. During the years 1998-2000 she worked in a project called EULEGIS, which was developing a unified interface for different legal databases in the Internet. Before that she worked in RASKE project, which developed means and methods for deployment of structured SGML documents in major Finnish public sector organisations. RASKE was a joint project, whose participants included Parliament of Finland, and ministries of the Finnish government.

Pasi T. Tiitinen

Researcher

University of Jyväskylä

Department of Computer Science and Information Systems

Finland

Email: pti@cc.jyu.fi

Pasi Tiitinen — Pasi Tiitinen has been doing research related to structured SGML/XML documents at the Department of Computer Science and Information Systems in the University of Jyväskylä since 1996. At the moment he works in a project called inSGML, which is developing, testing and customizing methods for the SGML/XML standardization process especially for industrial purposes. The work is related to his Doctoral Thesis, whose subject is usability of structured documents. During the years 1998-2000 he worked in a project called EULEGIS, which was developing a unified interface for different legal databases in the Internet. Before that he worked in RASKE project, which developed means and methods for deployment of structured SGML documents in major Finnish public sector organisations. RASKE was a joint project, whose participants included Parliament of Finland, and ministries of the Finnish government.

Airi Salminen

Professor

University of Waterloo

Department of Computer Science

Canada

Email: asalminen@db.uwaterloo.ca

Airi Salminen — Airi Salminen is a Professor in the Department of Computer Science and Information Systems at the University of Jyväskylä (Finland) and a Visiting Professor at the University of Waterloo (Canada). She received her Ph.D. in Computer Science from the University of Tampere in 1989. She was responsible for planning a new Master's Program in Digital Media at the University of Jyväskylä and has headed the program from its beginning in 1995. She has been the leader of several projects where research has been tied to document management development efforts in major Finnish companies or public

sector organizations. Her current research interests include document management, structured documents, XML, user interfaces, and software environments.